Department of Toxic Substances Control

Maureen F. Gorsen, Director
8800 Cal Center Drive
Sacramento, California 95826-3200

Linda S. Adams
Secretary for
Environmental Protection

Arnold Schwarzenegger
Governor

Arsenic Strategies
Determination of Arsenic Remediation Development of Arsenic Cleanup Goals For Proposed and Existing School Sites

March 21, 2007

The Department of Toxic Substances Control (DTSC) oversees the environmental assessments of proposed and existing school sites. During the Preliminary Environmental Assessment (PEA) or Remedial Investigation (RI) for school sites, arsenic may be identified as a chemical of concern based on comparisons to naturally occurring background concentrations. Once arsenic has been identified as a chemical of concern, a standard approach is needed to determine if remedial action is warranted and, if so, how to develop appropriate cleanup goals. The following is the suggested approach from the DTSC Human and Ecological Risk Division (HERD) for arsenic remediation on school sites.

Determination of Necessity for Remedial Action

Once arsenic concentrations have been identified to be above background levels, additional characterization may be required to determine the lateral and vertical extent of contamination. This information should be used in the decision making process for the necessity of a removal action. For the areas with elevated arsenic concentrations, if the data from the step out samples indicate that they are isolated areas (i.e., no real extent of contamination), no remedial action may be an option. For areas with high levels of arsenic concentrations, this approach may not be applicable. The complete data set for arsenic should be considered in the determination, including background, onsite ambient levels, and potential contamination.

Development of Cleanup Goals

The following are two options for developing a cleanup goal for arsenic.

Option 1

The upper limit of the background data set can be selected as the cleanup goal.

Option 2

Cleanup goals may be developed using the site specific data set for the school project. This data set may include both the data from the school site as well as background

values from the immediate area. The approach uses both visual evaluation of the data plots (graphical evaluation) and statistical calculations (statistical evaluation).

Graphical Evaluation

Step 1: Create normality plots. The plot should be completed using both log transformed and non-transformed data.

Step 2: For limited data sets, visually determine the inflection point in the distribution. This inflection point can in some cases be used as the approximation for a cleanup goal.

Statistical Evaluation

Step 1: After entering the data into an Excel spreadsheet, calculate summary statistics for the data set (e.g., mean, standard deviation, first quartile and third quartile). If the data set is sufficiently large, evaluate outliers in the data set. Suggested techniques include the ***fourth spread***, or other comparable techniques. Remove outliers from data set and estimate the Upper 95% Limit for the 0.99 Quartile $UL_{0.95}(X_{0.99})$ as described by Gilbert (1987).

Step 2: Recalculate summary statistics, including 95% Upper Confidence Limit (UCL) using modified data set.

Step 3 (optional): Comparisons of arsenic concentrations corresponding to the approximated inflection point with the summary statistics from data set excluding outliers.

Discussion of Uncertainties

- The incremental cancer risk difference between background levels and proposed cleanup goals will be very small or insignificant in most cases.

- Soil cleanup goals do not take into consideration potentially limited bioavailability of arsenic in soil. Most toxicology studies were based on arsenic in water, which is considerably more bioavailable.

Examples of Derivation of Arsenic Cleanup Goals

Example 1: Simple, Graphical Determination of the Arsenic Clean-up Goal for a School Site

The following example utilizes an actual data set from a school site in southern California. This example represents a rather large data set with 651 sample values. Table 1 summarizes the data set statistics.

Table 1
Arsenic Data Set Summary Statistics

DESCRIPTIVE STATISTIC	VALUE
Number of Samples	651
Minimum Concentration	0.27
Maximum Concentration	33
Mean Concentration	6.9
Median Concentration	6.7
Standard Deviation	4.02

Figure 1 presents the normality plot of the raw arsenic data. As can be seen from the plot, the data appears to be normally distributed and linear in the range from 1.0 up to about 12 mg/kg, where a distinct change in slope can be seen. This linear portion of the curve would be representative of ambient arsenic in this typical, urban environment. The inflection point where the slope changes is indicative of a population different from ambient arsenic, i.e., site contamination. For this example, 12 mg/kg would represent the upper-bound of ambient arsenic in soil at this site and would serve as the clean-up goal for arsenic.

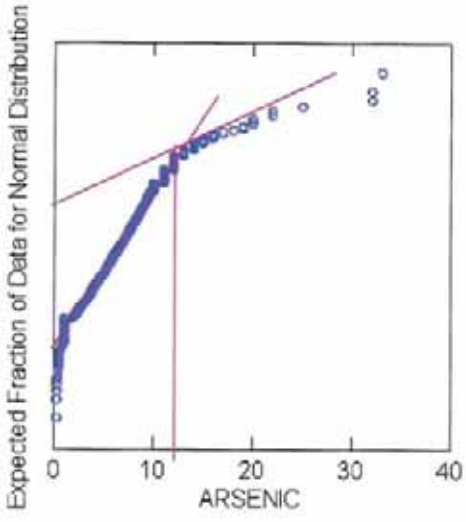

Figure 1
Normality Plot of the Arsenic Data Set

Example 2: Statistical Determination of the Arsenic Clean-up Goal for a School Site

The following example utilizes a combined data set made up of 19 individual school sites in southern California in order to exemplify the statistical determination of an arsenic clean-up goal. Figure 2 presents a plot of the frequency verses arsenic concentration, also known as a histogram. The shape of the histogram clearly demonstrates a classical, lognormal distribution. The descriptive statistics for the "Log-Transformed" combined arsenic data set of 1097 samples are summarized in Table 2.

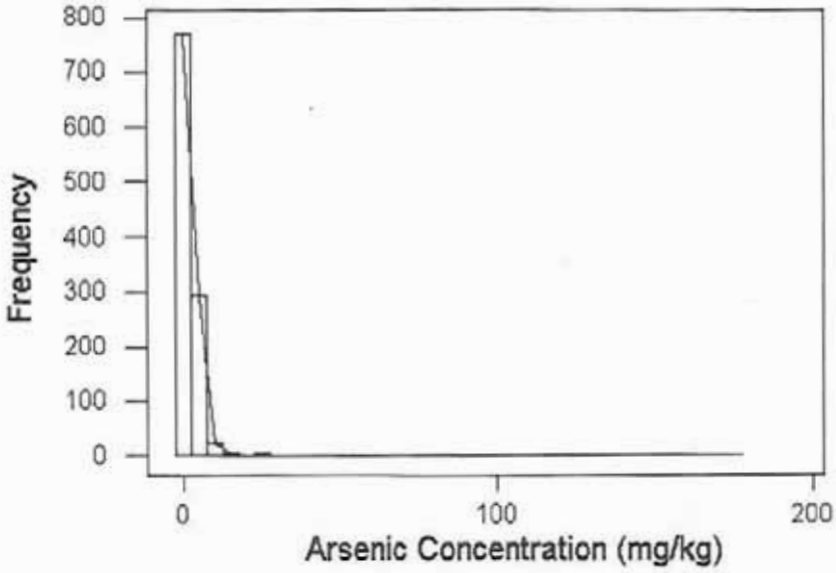

Figure 2
Histogram of the Arsenic Data

Table 2
Descriptive Statistics of the Combined Arsenic Data Set

DESCRIPTIVE STATISTIC	VALUE
Sample Size (n)	1097
Mean (μ)	0.1873 (1.54 mg/kg)
Median	0.1761 (1.50 mg/kg)
Standard Deviation	0.3916
Standard Error of the Mean[1]	0.0118
Minimum Concentration	-1.7620 (0.02 mg/kg)
Maximum Concentration	2.2480 (177 mg/kg)
Lower Quartile (Q_1)	-0.1249
Upper Quartile (Q_3)	0.4502

[1] The *Standard Error of the Mean* = $\frac{Std.Dev.}{\sqrt{n}}$

Because of the large sample size, wide range of arsenic concentrations and obvious extremes of the distribution, the data were analyzed for values that do not conform to the pattern established by the majority of values in the data set, e.g., ***outliers***. To determine the outliers in the arsenic data set, a pictorial summary called the box plot was utilized. A box plot describes the most prominent features of a data set, including 1)

center, 2) spread, 3) extent and nature of any departure from symmetry and 4) identification of any outliers or observations that lie unusually far from the main body of data. A box plot is based on measures that are unaffected by the presence of a few outliers, also known as the ***fourth spread***, f_s. The fourth spread, f_s, is defined as the measure of spread in a data set that is resistant to outliers and is calculated according to the following equation.

$$f_s = Q_3 - Q_1 \qquad \textbf{(Equation 1)}$$

$$= (0.4502 - (-0.1249))$$

$$= 0.5751$$

By definition, any observation farther than $1.5f_s$ from the closest fourth is considered an outlier. For the combined arsenic data set, $1.5f_s$ is equal to 0.8627 and any observation below $Q_1 - 1.5f_s$ or above $Q_3 + 1.5f_s$ would be considered an outlier. For the combined arsenic data set, outliers were defined as all observations:

$< Q1 - 1.5f_s$	and	$> Q3 + 1.5f_s$
$< (-0.1249 - 0.8627)$	and	$> (0.4502 + 0.8627)$
< -0.9876 (0.103 mg/kg)	and	> 1.3129 (20.55 mg/kg)

Therefore, the following arsenic concentrations were determined to be outliers: 177, 61.4, 49.2, 31.0, 27.6, 26.5, 24.0, 23.3, 22.7, 0.067 and 0.0173 mg/kg. These results are consistent with the box plot of the combined arsenic data set (Figure 3), which indicates that the nine largest and two lowest values are outliers.

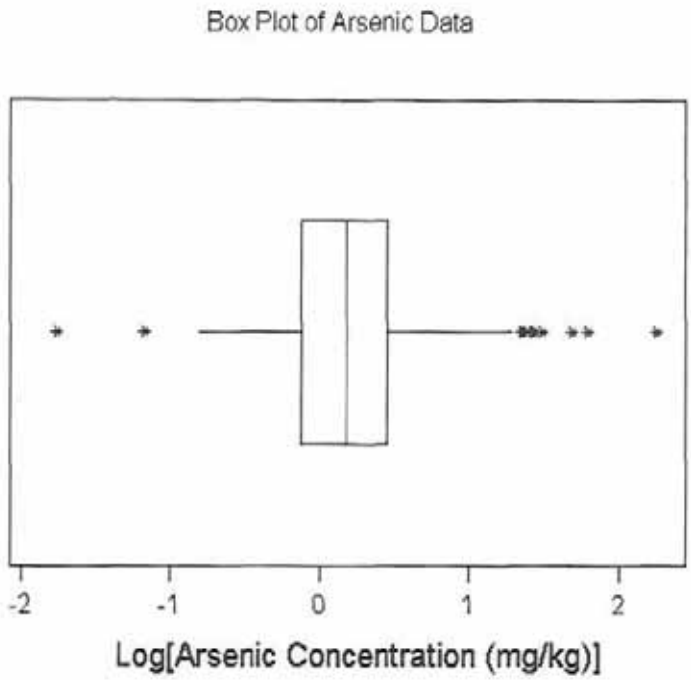

Figure 3

The large number of data points well characterizes the extremes of the distribution, thereby making it possible to use an estimate of an upper percentile of background concentrations as the value to be compared with the onsite C_{max}.

For this analysis, the 95% Upper Confidence Limit on the 99th-Percentile was chosen as the upper limit concentration. An upper 100(1 - α)% confidence limit for the true pth quantile, x_p, can be calculated if the underlying distribution is normal. As shown in Figure 4, the normal probability plot of arsenic data, excluding the outliers, clearly shows that the arsenic data is not normally distributed. However, as shown in Figure 5, the log-transformed arsenic data is normally distributed (i.e., the arsenic data fits a lognormal distribution). The descriptive statistics for the log-transformed arsenic data set, excluding the outliers previously established, are summarized in Table 3.

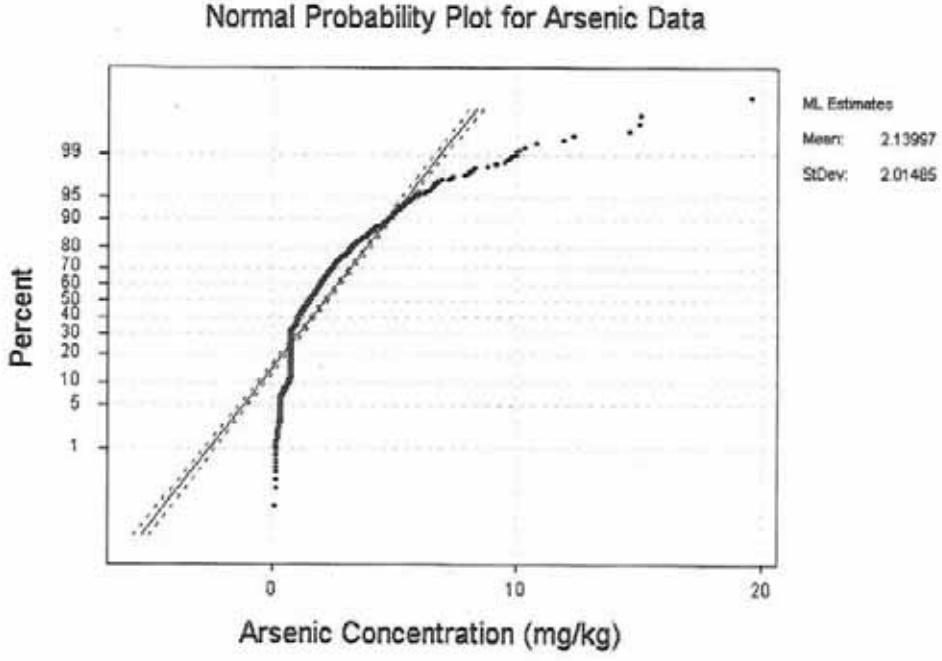

Figure 4

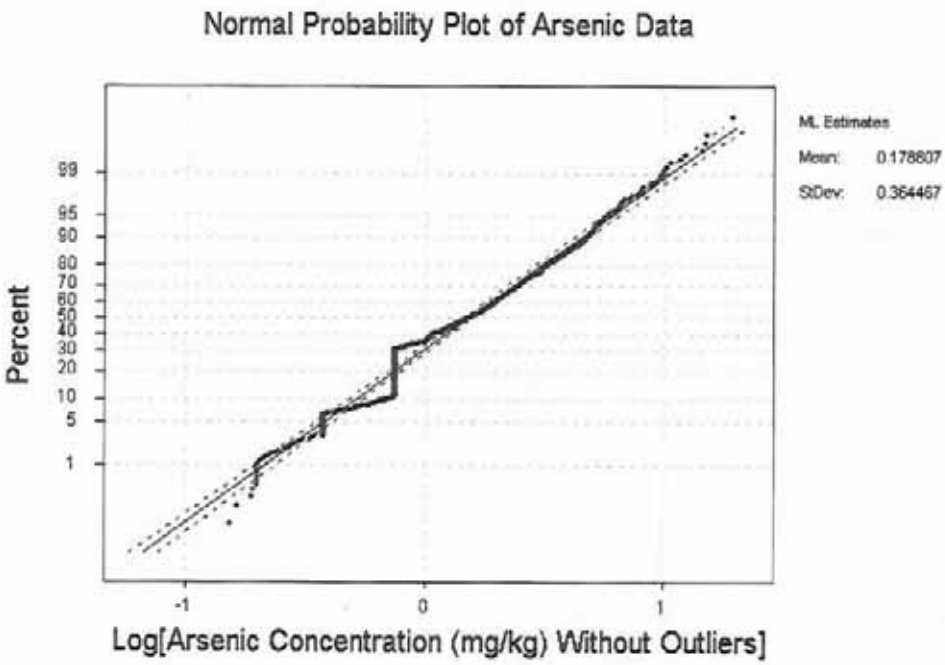

Figure 5

Table 3

Descriptive Statistics of the Combined Arsenic Data Set Without Outliers

DESCRIPTIVE STATISTIC	VALUE
Sample Size (n)	1086
Mean (μ)	0.1788 (1.51 mg/kg)
Median	0.1761 (1.50 mg/kg)
Standard Deviation	0.3646
Standard Error of the Mean[1]	0.0111
Minimum Concentration	-0.8125 (0.15 mg/kg)
Maximum Concentration	1.2930 (19.63 mg/kg)
Lower Quartile (Q_1)	-0.1249
Upper Quartile (Q_3)	0.4472

[1] The *Standard Error of the Mean* = $\frac{Std.Dev.}{\sqrt{n}}$

The upper limit of the data set can be estimated according to the following equation:

$$UL_{1-\alpha}(x_p) = \bar{x} + sK_{1-\alpha,p} \qquad \textbf{(Equation 2)}$$

Where,

$UL_{1-\alpha}(x_p) = The\ Upper\ Limit\ of\ the\ data\ set$

$\bar{x} = Mean\ \ of\ the\ data\ set$

$s = Std.\ Dev.\ of\ the\ mean$

$K_{1-\alpha,p} = Statistical\ tolerance\ factor\ \ for\ estimating\ an\ Upper\ 100(1-\alpha)$
$Confidence\ Limit\ on\ the\ pth\ Quantile$

For calculating the 95% confidence limit of the 99th quantile of the arsenic data set, excluding outliers, $K_{0.95,\ 0.99}$ **= 2.40** (from Table A3, Gilbert 1987). Using the mean and standard deviation of the arsenic data set (Table 2), the $UL_{0.95}(X_{0.99})$ can be calculated as follows:

$$UL_{0.95}(X_{0.99}) = 0.1788 + (2.40)(0.3646)$$
$$= 1.054$$

Since the arsenic data is log-transformed, the Upper Limit Concentration is the antilogarithm of this value.

$$UL_{0.95}(X_{0.99}) = 10^{1.054}$$

= 11.32 mg/kg

A distribution-free, non-parametric analysis was also conducted to estimate the theoretical ***$UL_{0.95}(X_{0.99})$*** as described by Gilbert (1987). This method, also known as the distribution-free technique, is used when the underlying distribution is either unknown or non-normal. This method was employed using the following equation:

$$The\ Rank\ of\ the\ UL_{0.95}(X_{0.99}) = p(n+1) + Z_{1-\alpha}[np(1-p)]^{1/2}$$ **(Equation 3)**

Where,

$$p = 99th Quantile = 0.99$$
$$Z_{1-\alpha} = Z\ Value\ for\ the\ 95\%\ Confidence\ Interval$$
$$= Z_{0.95}$$
$$= 1.645$$
$$n = Number\ of\ samples, excluding\ outliers$$
$$= 1086$$

For the arsenic data set, the Rank of the Upper 95% Limit for the 0.99 Quantile (***Rank of $UL_{0.95}(X_{0.99})$***) can be calculated as follows:

$$Rank\ of\ UL_{0.95}(X_{0.99}) = 0.99(1087) + 1.645[1086(0.99)(0.01)]^{1/2}$$
$$= 1081.524$$

Then, the ***$UL_{0.95}(X_{0.99})$*** would be the arsenic concentration that is 52.4% of the way between the 1081st and the 1082nd largest values. Since the 1081st value is 11.9 mg/kg and the 1082nd value is 12.3 mg/kg, the ***$UL_{0.95}(X_{0.99})$*** would be approximately **12 mg/kg**.

Example 3: Determination of the Arsenic Clean-up Goal for a School Site

Examples 1 and 2 represent very large, ideal arsenic data sets used to demonstrate the graphical and statistical approaches to setting clean-up goals. The following example utilizes a much smaller and typical arsenic data set from a school site in Southern California and demonstrates several methods for determination of arsenic cleanup goals.

Method 1. Graphical Evaluation

Step 1. Graphical representations of arsenic data set.

Create Normality plots using both raw and log transformed data as shown in Figures 6 and 7. The arsenic concentration can be plotted as a function of the expected value for a normal distribution or alternatively, the data set can be plotted from least value to highest value as the cumulative percent of samples. Either graphical treatment results in a curve representing the distribution of the data set.

Step 2. Visual inspection of the curves

Visual inspection of the curve may yield a determination of an inflection point which represents a break between the ambient level of arsenic for the site and the portion of the curve that represents a separate, higher population which may be a consequence of a release to the environment. For the example shown below it can be determined that an inflection point in the distribution of samples occurs at an approximate arsenic concentration of 10 mg/kg (Figure 6) or at the Log_{10}[arsenic concentration] value of 1 which corresponds to 10 mg/kg (Figure 7).

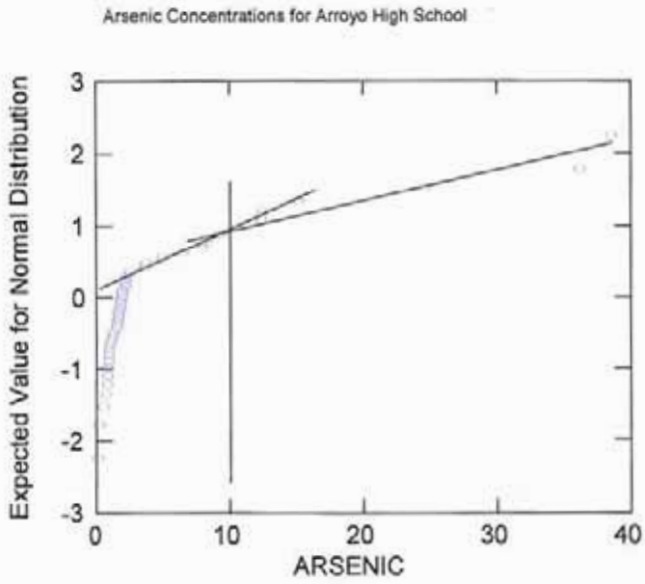

Figure 6
Distribution of arsenic concentrations in mg/kg

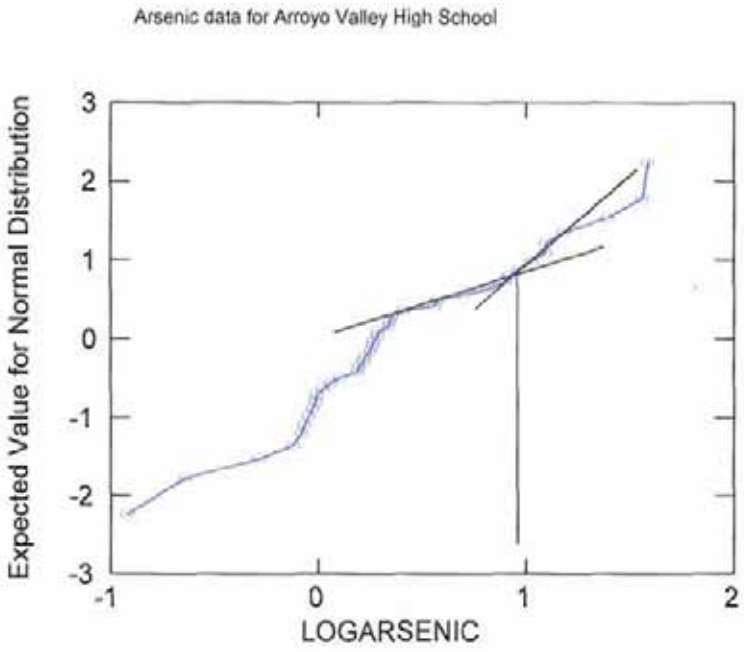

Figure 7
Distribution of arsenic in a log transformed representation

Method 2. Statistical Evaluation: Quartile Analysis ("Fourth Spread")

A statistical approach may be used that identifies upper-bound outliers which can then be removed from the data set to generate a new data set for which an upper confidence limit (UCL) can be defined and utilized as the cleanup goal.

Step 1. Derivation of Descriptive Statistics:

Descriptive statistics as shown in Table 4 were calculated for this site based on the site-specific arsenic data set. These statistics included: number of samples, minimum and maximum site concentration, mean, standard deviation, sample distribution, median and quartiles, 95th and 98th percentile and 95% UCL.

Table 4
Descriptive Statistics

DESCRIPTIVE STATISTIC	VALUE
Number of samples	40
Minimum detected value	0.12
Maximum detected value	38.6
Mean	5.75
First quartile (Q1)	0.98
Median	1.85
Third quartile (Q3)	4.98
95^{th} percentile	25.18
98^{th} percentile	36.73
95%UCL of mean	8.61
Standard deviation	8.93

Values listed in mg/kg

Step 2. Determine upper-bound outliers:

The quartile analysis to determine upper-bound outliers in the data set may be conducted as in the following example: The median and first and third quartiles from the data set shown in Table 4 were determined and a fourth spread (*Fs*) was generated.

First quartile (Q1) = 0.98
Median (second quartile, Q2) = 1.89
Third quartile (Q3) = 4.98

Fs = (Q3 –Q1) = 4.0

Outliers for the upper bound of the site-specific arsenic concentrations are defined as:

All data points greater than Q3 + [1.5 x *Fs*]: 4.98 + 6.0 = 10.98.

Therefore, any value higher than 10.98 mg/kg is considered an outlier (contaminated soil sample) and is eliminated from the data set because it is higher than the ambient level.

Step 3. Statistical re-evaluation of the data set.

The site-specific data set is then re-evaluated with outliers removed to create the adjusted site ambient data set. The statistical evaluation of the adjusted ambient data set yields the following values:

Table 5
Arsenic data set statistics with upper-bound outliers removed

Number of samples	35
Minimum detected value	0.12
Maximum detected value	10.6
Mean	3.74
Std deviation	6.49
98th percentile	9.72

Values listed in mg/kg

An appropriate cleanup goal for this site is the 98th percentile of the adjusted arsenic data set, which is approximately 10 mg/kg. Note that the 98th-percentile was used as an upper-bound for this data set due to the smaller number of samples (N = 40).

References

Gilbert, Richard O. 1987: Statistical Methods for Environmental Pollution Monitoring. Van Norstrand Reinhold Company, Inc.